UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

Benign Overfitting and its Applications in Finance

Author : Tommaso Cassio
Supervisor : Nathan Lassance
Academic Year 2023-2024
Business Engineering – Financial Engineering

Declaration

During the preparation of this master's thesis, the author utilized ChatGPT, Grammarly for the following purpose:

1.Confrontation: ChatGPT was principally used to verify my understanding of the scientific literature and ask questions about the accuracy of my interpretations. ChatGPT was also used to manage errors during the implementation of Python code for the empirical section.

2. Layout: ChatGPT was used for rewriting more rapidly tables in LaTeX.

3. Rewriting: Grammarly, a grammar checker and rewriting assistant, helped me to write this master's thesis to make it more academic.

After using ChatGPT, Grammarly, the author diligently reviewed and edited the content produced by the tool. We take full responsibility for the final content presented in this thesis.

By signing this declaration, we affirm that the content of this master's thesis reflects our original work, augmented by the responsible use of AI.

Signed by: Tommaso Cassio 29th July 2024

Summary

Benign overfitting is a complex and recently discovered phenomenon that remains not fully understood. It describes scenarios where a model fits perfectly the training data and performs well on unseen testing data. This phenomenon contradicts the traditional belief that overfitting normally results in poor data generalisation. In this master's thesis, we will explore the current literature on this theory and apply these insights to finance.

Firstly, we tested some benign overfitting conditions on multiple linear regression and ridge regression models. We evaluated the presence of possible benign overfitting by examining the excess risk and utilizing a pre-established taxonomy. Secondly, we introduced the concept of the double descent curve, an extension of the bias-variance trade-off curve in scenarios of overfitted models. We analysed this phenomenon in regression models and we divided the risk into its variance and bias components to understand their respective influences. Lastly, we investigated overfitting in modern portfolio theory by comparing four portfolio models across various performance metrics. We also assessed the portfolios' variance as the number of variables increased.

Our findings indicate that most of the benign overfitting conditions on our datasets were met. The regression models also demonstrated a small excess risk, suggesting possible benign overfitting. The double descent curve was evident, showing reduced risk when the model was overfitted and reinforcing our hypothesis of benign overfitting. Furthermore, the double descent curve is mostly due to increased variance and bias at a certain point. We also observed that shrinkage can be beneficial in certain situations to reduce risk even in overfitted models. However, for modern portfolio theory, we could not conclusively determine the presence of benign overfitting in any model. We only noted that portfolio variance decreased when the number of assets exceeded the number of observations.

In conclusion, we observed that benign overfitting can be applied to financial datasets, but its applicability depends on both the dataset and the model. Even if the conditions are met, the model's role is crucial in determining the presence of benign overfitting. It is also possible that benign overfitting conditions may not be extended beyond regression models.

Acknowledgements

I want to begin by acknowledging my supervisor, Professor Nathan Lassance. Your availability and advice during our meetings have been invaluable. Your expertise has guided me through numerous discussions, helping me to address this master's thesis and interpret my results effectively. For this, I am deeply grateful.

I also want to thank my parents. Your support and belief in me have been the driving force for my motivation and success. To my brothers, thank you for always being there, for your constant support and understanding. Your confidence in me has inspired me to always aim higher. Lastly, I want to thank my entire family in Italy for their support throughout these years.

To my beloved, thank you for your endless encouragement and kindness, especially during our time apart while I was in Dublin and during my internship. Your support has been a massive strength.

Thank you to my friends for the past five university years. Our friendship has been a source of joy and a necessary distraction during the intense period of writing this thesis. The countless memories we have made and the laughter we shared. Your support and the moments of relaxation you offered were essential in keeping me motivated.

Finally, I would like to extend my gratitude to the UCLouvain and the Louvain School of Management. The interesting programs you offer have helped me acquire various skills and knowledge. The experiences and lessons I have learned here will stay with me throughout my life.

Contents

List of Tables

List of Figures

Introduction

Benign overfitting is a complex phenomenon that is not yet fully understood. It refers to situations where a model fits the "training" data perfectly, including noise, and still performs well on unseen "testing" data. This discovery is contrary to traditional belief where overfitting typically leads to poor generalisation of new data. Several studies have already been conducted to understand this benign overfitting phenomenon, though most of the research focused on complex models used for image classification.

To advance the research of this phenomenon in finance, the primary objective of this master's thesis is to investigate the potential occurrence of benign overfitting in financial data. If this phenomenon occurs, we aim to determine whether it is influenced by the dataset, the model, or both. To analyse this, we will use Kenneth R. French datasets and start with basic models of multiple linear regression and ridge regression. Subsequently, we will investigate whether benign overfitting can be applied to modern portfolio theory.

This master's thesis starts with an examination of the hypotheses and assumptions under which benign overfitting can occur. We will discuss the conditions necessary for benign overfitting in regression models, explained by Bartlett et al. (2020); Tsigler and Bartlett (2023). Then, we will explain the concept of the double descent curve described by Li, Su, and Sejdinovic (2021); Mikhail Belkin et al. (2019) and provide an overview of the overfitting taxonomy detailed by Mallinar et al. (2022). After that, we will explain the importance of model specification and the relevance of the double descent curve in financial data, referencing studies by Kelly and Xiu (2023); Kelly et al. (2021); Kelly et al. (2022). Lastly, to conclude the theoretical part, we will detail several portfolio optimization models, with particular attention to the model proposed by Bodnar et al. (2023).

In the subsequent section, we will present our findings, which include the identification of benign overfitting in financial datasets utilizing both multiple linear regression and ridge regression models and the observation of the double descent curve. Additionally, we will compare the predefined portfolios using different metrics. Finally, this master's thesis will conclude with a summary of our findings.

Chapter I

Literature Review & Research Question

In this chapter, we will delve into the theoretical and previously established findings to gain a comprehensive understanding of benign overfitting. By explaining these theoretical concepts and existing research, we will be better positioned to understand the findings presented in this master's thesis.

1 Benign Overfitting in Regression Models

1.1 Linear Regression and Benign Overfitting

The easiest way to observe benign overfitting is in linear regression, as explained by Bartlett et al. (2020). The paper proposes specific conditions under which this phenomenon can be observed. However, to fully understand the findings of this paper, we should first define some statistical concepts.

Linear Regression

Linear regression is one of the most fundamental methods for supervised statistical learning (James et al., 2023). The primary goal of these models is to predict an outcome y under a distribution shift from other variables X (Hastie et al., 2009); (Mallinar et al., 2024). To achieve this, we start with a dataset containing n observations and p variables. Typically, the dataset will be divided into two distinct sections. The first part, normally large, is known as the training dataset and is used to train the model. The second part, the testing dataset, is used to evaluate the model's accuracy. We assume an independent and identically distributed sample, denoted as $(x^i, y^i)_{i=1}^n \sim \mathcal{D}_s^n$. Where s represents the "source" (training dataset). We define the distribution $\mathcal{D}_s$ for the training dataset. We also determine t which stands for the "test" dataset and $\mathcal{D}_t$ represents the testing dataset distribution.

Then, we form a data matrix $X \in \mathbb{R}^{n \times p}$ and a corresponding outcome vector $y \in \mathbb{R}^n$. The linear regression model is defined as follows:

$$y = X\theta^* + \varepsilon,$$

The noise vector ε are independent and identically distributed, each having a mean of zero and variance v_ε^2. The vector θ^* represents the optimal parameters used to predict the outcome y. The covariance matrices are denoted by $\Sigma \in \mathbb{R}^{p \times p}$.

Bartlett et al. (2020) provides a more precise definition of linear regression:

1. The covariance matrix $\Sigma = \mathbb{E}[xx']$.
2. The optimal parameter vector $\theta^* \in \mathbb{R}^p$ satisfies

$$\mathbb{E}[(y - x'\theta^*)^2] = \min_\theta \mathbb{E}[(y - x'\theta)^2]$$

   This implies that θ^* must be the optimal linear predictor for minimising the mean squared error.

Bartlett et al. (2020) also takes some assumptions defined as follows:

1. x and y are mean zero.
2. The covariate x can be expressed as:

$$x = V\Lambda^{1/2}z$$

   where $\Sigma = V\Lambda V'$ is the spectral decomposition of Σ, V is an orthogonal matrix of eigenvectors, Λ is a diagonal matrix of eigenvalues and z is a vector with independent σ_x^2-subgaussian components. Subgaussian means that z has tail behaviour similar to a Gaussian distribution, with exponentially decaying tails making extreme values unlikely.
3. The conditional noise variance should be bounded below by some constant σ^2:

$$\mathbb{E}[(y - x'\theta^*)^2|x] \geq \sigma^2.$$

4. The noise $y - x'\theta^*$ is σ_y^2-subgaussian, conditionally on x. That is, for all $\gamma \in \mathbb{R}$,

$$\mathbb{E}[\exp(\gamma(y - x'\theta^*))|x] \leq \exp(\sigma_y^2\gamma^2/2)$$

This also implies that the conditional expectation of y given x is:

$$E[y \mid x] = x'\theta^*.$$

Minimum Norm Estimator

The optimal parameter θ^* defined above is not directly given and must be calculated. We use $\|.\|$ to denote the Euclidean norm of a vector in $\mathbb{R}^n$. This norm, also called ℓ_2, can be calculated as the square root of the sum of squares of its components:

$$\|\theta\| = \sqrt{\sum_{i=1}^{n} \theta_i^2}$$

Bartlett et al. (2020) uses this formula to determine the minimum norm estimator, which is the parameter θ with the smallest squared sum norm:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta}\{\|\theta\|^2 : X'X\theta = X'y\} \\ &= X'(XX')^{-1}y.\end{aligned}$$

At this point, we consider XX' invertible because we assume that $n > p$. When p is larger than n, we use the pseudo-inverse matrix $(XX')^+$.

Effective Rank

The effective rank is an important condition for benign overfitting. This one is denoted $r_k(\Sigma)$ and $R_k(\Sigma)$, and measures the effective dimensionality of the data space based on the sum of the eigenvalues of Σ. It helps determine the noise distribution across the parameter space and its impact on prediction accuracy. The parameter k represents the number of larger eigenvalues in the linear regression model. A large k indicates a more complex capability of capturing relationships between the independent variables X and the dependent variable y.

The covariance operator Σ, is composed of the eigenvalues λ_i, where $i = 1, 2, \ldots, n$.

If $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$ for $k \geq 0$, we define:

$$r_k(\Sigma) = \frac{\sum_{i>k}^{k} \lambda_i}{\lambda_{k+1}}, \quad R_k(\Sigma) = \frac{\left(\sum_{i>k}^{k} \lambda_i\right)^2}{\sum_{i>k}^{k} \lambda_i^2}.$$

The effective rank $r_k(\Sigma)$ is based on a weighted sum of the eigenvalues. A smaller $r_k(\Sigma)$ suggests that the covariance matrix has fewer important directions for prediction. In other words, it estimates the amount of variance explained by the eigenvalues smaller than the chosen k-th eigenvalue. On the other hand, $R_k(\Sigma)$ reflects the rate at which the eigenvalues of the covariance matrix decrease. A large $R_k(\Sigma)$ indicates that a few eigenvalues dominate the variance, whereas a smaller one tells us that the eigenvalues will be more uniformly distributed.

Excess Risk Estimator

The most important factor in determining benign overfitting is the excess risk, defined as the additional risk that an estimator θ has compared to the optimal estimator θ^*. The excess risk, denoted as $R(\theta)$, is mathematically expressed as the increase in mean squared error (MSE) between θ^* and θ. The excess risk explained by Bartlett et al. (2020) is defined as:

$$R(\theta) = \mathbb{E}_{x,y}\left[(y - x'\theta)^2 - (y - x'\theta^*)^2\right]$$

where $\mathbb{E}_{x,y}$ is the conditional expectation given all random quantities other than x, y.

The main criterion for evaluating benign overfitting is that the excess risk should be small. However, Bartlett et al. (2020) does not specify a particular threshold for $R(\theta)$.

Conditions for Benign Overfitting in Linear Regression

Relation between p and n Bartlett et al. (2020) emphasizes that to be sure that benign overfitting occurs, the number of variables p in the model should exceed the number of observations n. This condition ensures that the model has sufficient flexibility to capture the variability in the data without being overly constrained by the limited number of observations. This relationship can be expressed as $p > n$, or as $c > 1$, where c is the ratio between the number of variables p and the number of observations n.

Effective Rank Condition The second condition defined by Bartlett et al. (2020) to obtain benign overfitting is that the effective rank of Σ, $R_k(\Sigma)$, should be large compared to the sample size n. This can be expressed as $R_k(\Sigma) \gg n$ and $\lim_{n\to\infty} \frac{n}{R_k} = 0$. Bartlett et al. (2020) states that this condition is necessary for significant overfitting which could allow nearly optimal prediction.

Eigenvalues Conditions The eigenvalues of the covariance matrix Σ provide insights into the variance distribution in the data. Bartlett et al. (2020) identifies specific characteristics of these eigenvalues associated with benign overfitting.

1. The first condition of eigenvalues includes that there should be a dominant first eigenvalue (λ_1), indicating a strong signal in the data, as well as a substantial number of non-zero eigenvalues relative to the number of parameters p: $\lambda_1 \gg \lambda_2 \geq \ldots \geq \lambda_p$. The presence of many eigenvalues smaller than the leading one further indicates a smooth variance distribution.

2. The last important condition is that the cumulative sum of eigenvalues should be smaller than n because it suggests that the scale of the problem is low compared to the number of observations. This can be written as $\sum_{i=1}^{p} \lambda_i \ll n$.

These four conditions are important for assessing whether the data used can be subject to benign overfitting in linear regression and whether the excess risk could be low.

1.2 Benign Overfitting in Ridge Regression

To delve deeper into the theory of benign overfitting, we will examine overfitting in ridge regression. Ridge regression is typically used to avoid overfitting. However, Tsigler and Bartlett (2023) explains that it can be useful in benign overfitting. The paper by Tsigler and Bartlett (2023) is built on the conclusions presented by Bartlett et al. (2020), where the number of variables p exceeds the number of observations n.

The regression model remains the same, where the optimal estimator $\theta^* \in \mathbb{R}^p$ is used, and the relationship is given by:

$$y = X\theta^* + \varepsilon,$$

where ε is the noise vector, consisting of i.i.d. centred random variables with variance v_ε^2.

Ridge Regularisation

Ridge regression is similar to multiple linear regression but primarily aims at avoiding overfitting by adding a regularisation parameter λ (Montesinos López et al., 2022). This parameter is introduced in the minimum norm estimator that estimates $\hat{\theta}$ from X and y:

$$\hat{\theta} = X'(XX' + \lambda I_n)^{-1} y$$

where I_n is the identity matrix of size $n \times n$.

The matrix $(\lambda I_n + XX')$ is called the Gram matrix A (Tsigler & Bartlett, 2023).

$$A = \lambda I_n + XX'$$

When $\lambda = 0$, also called the ridgeless case, we obtain the same multiple linear regression formula discussed in Section 1.1. Ridge regularisation shifts all eigenvalues by λ. In ridgeless regression, if the original eigenvalues of A were $\lambda_1, \lambda_2, \ldots, \lambda_n$, the eigenvalues of the new regularised matrix will be $\lambda_1 + \lambda, \lambda_2 + \lambda, \ldots, \lambda_n + \lambda$.

Another important recall from Bartlett et al. (2020) is the effective rank. Tsigler and Bartlett (2023) use a more general definition that includes ridge regularisation:

$$r_k(\Sigma) = \frac{1}{\lambda_{k+1}} \left(\lambda + \sum_{i>k}^{k} \lambda_i \right)$$

So, if we set $\lambda = 0$, $r_k(\Sigma)$ will have the same formula as defined earlier in Section 1.1. Additionally, Tsigler and Bartlett (2023) introduces another parameter $\rho_k(\Sigma)$, which measures how large the effective rank is compared to the number of observations:

$$\rho_k(\Sigma) = \frac{r_k(\Sigma)}{n}$$

Noise Distribution, Variance, and Bias

In the study of benign overfitting in ridge regression, Tsigler and Bartlett (2023) defines the bias (B) and variance (V) of a model as follows:

$$B = \|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2 = \|(X'A^{-1}X - I_p)\theta^*\|_\Sigma^2,$$

$$V = \mathbb{E}_\varepsilon \left[\frac{\|\hat{\theta}(\varepsilon)\|_\Sigma^2}{v_\varepsilon^2} \right] = \mathrm{tr}(A^{-1}X\Sigma X'A^{-1})$$

The bias and variance formulas show that these quantities depend only on the model and the data structure, not on the specific noise characteristics. This means that, regardless of how the noise in the data is distributed, these measures will consistently reflect the model's performance.

The main objective of Tsigler and Bartlett (2023) is to provide precise bounds on bias and variance that are valid for any sample size, not just in the limit as $n \to \infty$. These bounds are important because they help to understand the conditions under which benign overfitting can occur. Tsigler and Bartlett (2023) defines that the bias should be low enough

to capture the true signal without underfitting, while the variance should be controlled to prevent the model from fitting the noise too closely. These conditions ensure that the ridge regression model can perform well even in overparameterized settings, allowing benign overfitting where the model interpolates the training data but still maintains low prediction error on new data.

Regularisation and Benign Overfitting

Tsigler and Bartlett (2023) provides precise bounds on the bias term in ridge regression with a non-zero λ, extending the results of Bartlett et al. (2020). They replace the assumption of independence of components with a broader sufficient condition. This control implies that ρ_k is at least a constant, offering a more general condition than independence and a high effective rank. The key discovery from Bartlett et al. (2020) is the use of effective rank to evaluate benign overfitting, which Tsigler and Bartlett (2023) builds upon by separating the first k eigenvectors to bound the bias term.

Tsigler and Bartlett (2023) also introduces the concepts of "essentially low-dimensional" (spiked part) and "essentially high-dimensional" (tail part). The spiked part refers to the first k dimensions $(0 : k)$ with high eigenvalues, important for determining the bias in ridge regression with a regularisation coefficient of $\lambda + \sum_{i>k} \lambda_i$. The tail part $(k : \infty)$ includes smaller eigenvalues, where the signal is primarily embedded in the error.

Moreover, Tsigler and Bartlett (2023) demonstrates that if the noise and signal energy in the tail $(k : \infty)$ are small compared to the spiked part $(0 : k)$ and the effective rank r_k of the tail exceeds n, a negative λ might be necessary.

1.3 Double Descent Curve

The double descent curve presents a new understanding of the trade-off between prediction accuracy and model complexity, moving beyond the traditional bias-variance trade-off. Traditionally, machine learning models aim to balance between being too simple (underfitting) and too complex (overfitting). The classical bias-variance trade-off curve shows that when the model complexity increases, bias decreases and variance increases, with the optimal model complexity lying at the "sweet spot" that minimises test error. This curve can be observed in Figure I.1.

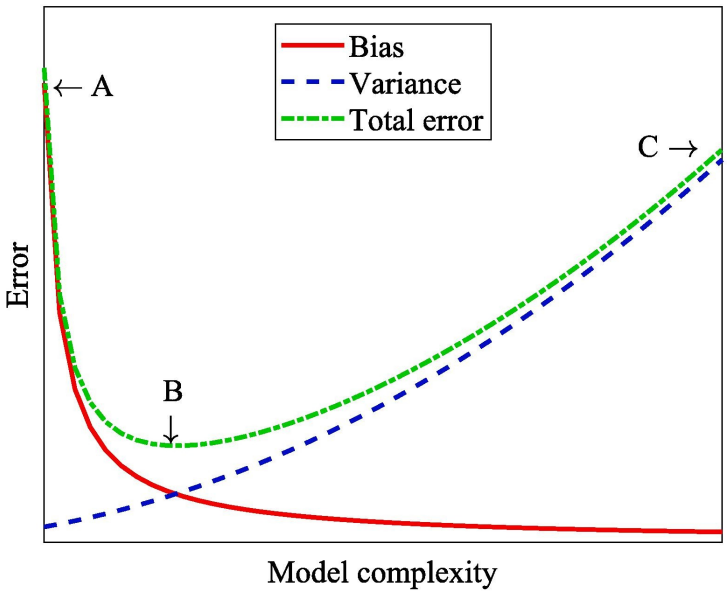


Figure I.1: Bias-variance trade-off curve. Where **B** represents the "sweet spot", (Guan & Burton, 2022, p. 19).

However, high-dimensional machine-learning practices challenge this traditional understanding by demonstrating that overfitted models can perfectly fit training data, including noise, and still perform well on test data. This phenomenon, known as benign overfitting, calls into question the applicability of the classical bias-variance trade-off in high-dimensional contexts.

In the paper by Mikhail Belkin et al. (2019), the authors introduce a performance curve called the "double descent" curve, which is an extension of the classical risk curve. Mikhail Belkin et al. (2019) demonstrates that at the interpolation threshold ($p = n$), the training risk achieves a value of zero, while the test risk spikes. After this threshold, the test risk declines and often reaches lower levels than at the classical "sweet spot", as you can see in Figure I.2. This double descent curve reveals that increasing model capacity beyond the interpolation point can improve performance, contrary to the conventional expectation that models achieving zero training error would generalise poorly. This suggests that large-dimensional models can yield lower risk than the sweet spot. However, it is important to note that this phenomenon is not always observed, and in some cases, the risk after the interpolation point can be higher than at the sweet spot. Various forms of bias-variance curves are presented in the paper by Yang et al. (2020), indicating that the double descent curve is not the only risk curve existing.

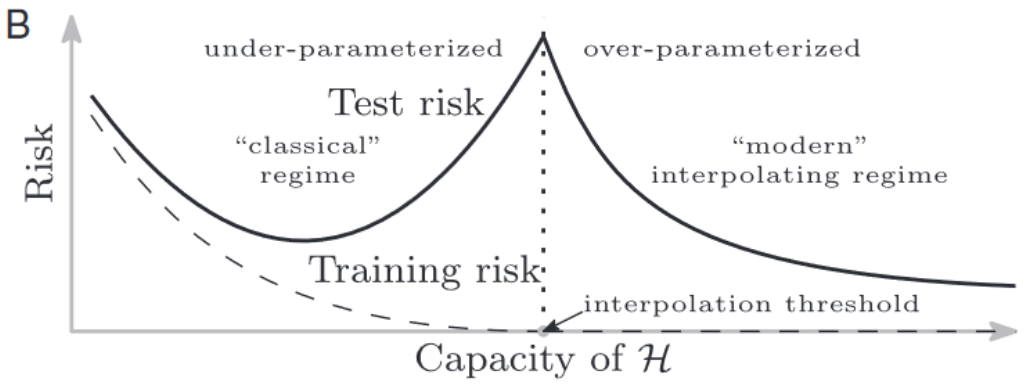


Figure I.2: Double-descent risk curve (Mikhail Belkin et al., 2019, p. 15850).

Similarly, Li, Su, and Sejdinovic (2021) investigated the double descent curve and show that at the interpolation threshold ($p = n$), the risk peaks and then decreases until around $p = n^2$, where it starts to rise again. Their analysis highlights the necessity of overparameterization to achieve the double descent curve. This phenomenon suggests that larger function classes can create simpler interpolating functions that generalise more effectively to unseen data. Furthermore, the findings of Li, Su, and Sejdinovic (2021) add that noise in the features can improve generalisation by acting as a regularizer.

This phenomenon of double descent curve has also been observed in more complex models (Muthukumar et al., 2020) and in models containing artificial neuronal networks (Belkin et al., 2019); (Yang et al., 2020).

1.4 Taxonomy

This section discusses the classification of benign overfitting and its occurrence in various machine learning and high-dimensional models. The authors of Mallinar et al. (2022); Mallinar et al. (2024) created a taxonomy for defining overfitting by classifying different models as benign or catastrophic and introduced the concept of tempered overfitting, which lies between the two. The different classifications can be observed in Figure I.3. In their research, Mallinar et al. (2022) classify different complex models using various image datasets. They discovered that ridgeless and ridge models exhibit tempered overfitting, a view that diverges from the conclusions in Tsigler and Bartlett (2023), which suggests that ridge regression is generally benign. However, this difference may arise from different experimental settings, as prior studies on benign overfitting, such as those cited in Bartlett et al. (2020) and Tsigler and Bartlett (2023), typically considered scenarios where the sample size n is large but still less than the dimensionality of p. The authors of the papers Mallinar et al. (2022); Mallinar et al. (2024) do not impose such restrictions. The regression models also have some differences between those articles.

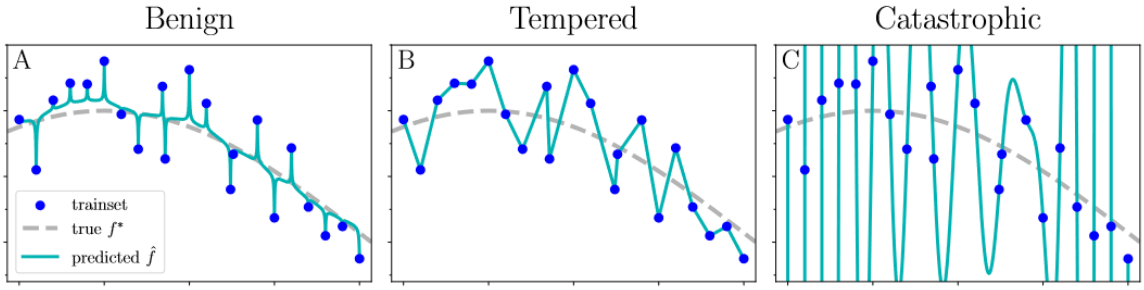


Figure I.3: **A.** Benign overfitting, the predictor approaches the optimal function. **B.** Tempered overfitting, the predictor approaches a constant test risk greater than the optimal risk. **C.** Catastrophic overfitting, the predictor generalises poorly. (Mallinar et al., 2022, p. 2)

However, Mallinar et al. (2022) noted a decrease in the excess risk associated with certain interpolating methods as the dimensionality of the space increases. This phenomenon, referred to as "blessing of dimensionality," is also indicated in works such as Belkin, Hsu, and Mitra (2018) and Rakhlin and Zhai (2019), suggesting that problems in higher dimensions tend to experience lower risks. To categorize the different types of overfitting, Mallinar et al. (2022) decided to use the excess risk. They defined that, as the number of samples approaches infinity, the risk behaves in three different ways. This taxonomy is outlined in Table I.1, where R is the excess risk $R(\theta)$ defined in Section 1.1. R^*, the optimal excess risk is normally close to zero.

Overfitting Classification	$R(\theta)$ **Behaviour**
Benign	$\lim_{n\to\infty} R_n = R^*$
Tempered	$\lim_{n\to\infty} R_n \in (R^*, \infty)$
Catastrophic	$\lim_{n\to\infty} R_n = \infty$

Table I.1: Overfitting Taxonomy (Mallinar et al., 2022, p. 5).

2 Benign Overfitting on Return Prediction

2.1 Introduction to Machine Learning Return Prediction Model

One of the first financial applications of overfitted models, which can potentially be benign, has been in return prediction models. Some papers have documented significant gains in performance through the use of machine learning. However, there remains a limited theoretical understanding of return forecasts and portfolios formed from these highly parameterized models. Kelly et al. (2021) investigates the performance of flexible non-linear return prediction models in regimes of high dimensionality and complexity.

Kelly et al. (2021) and Kelly and Xiu (2023) start by imagining a return prediction model. This model comes from a Data Generating Process (DGP), where the distributions, relationships, and rules govern the generation of data points. Here, this DGP is the asset return R that is generated by:

$$R_{t+1} = f(X_t) + \varepsilon_{t+1}$$

where the variable X may be known, but the function f is unknown. Since f is unknown, Kelly and Xiu (2023) states that they decided to approximate f with a basic neural network:

$$f(X_t) \approx \sum_{i=1}^{p} S_{i,t}\beta_i$$

where $S_{i,t} = \tilde{f}(w_i' X_t)$ is a known nonlinear activation function with known weights w_i and where p is large. An activation function is used to introduce non-linearity into a model. It allows the neural network model to create complex representations and functions based on the inputs, which is impossible with a simple linear regression model. When the data we wish to model is non-linear, we must account for this in our model. Non-linear activation functions are essential for handling non-linear problems such as classification, image recognition, and language processing (Kelly et al., 2022). The returns R can be determined as:

$$R_{t+1} = \sum_{i=1}^{p} S_{i,t}\beta_i + \tilde{\varepsilon}_{t+1}$$

Next, we should define a returns performance metric. A commonly used metric is the Sharpe Ratio, which involves subtracting the risk-free rate from the expected return and then dividing it by the standard deviation of the returns.

$$SR = \frac{R_t^{\pi} - R_f}{\sigma_t}$$

However, in Kelly et al. (2021), the authors aim mostly to optimise the unconditional Sharpe Ratio, defined as:

$$\widetilde{SR} = \frac{\mathbb{E}[R_{t+1}^{\pi}]}{\sqrt{\mathbb{E}[(R_{t+1}^{\pi})^2]}}$$

In this unconditional Sharpe Ratio definition, the expected return is divided directly by the root mean square of the returns. This version of the Sharpe Ratio is sometimes referred to as the "uncentered" Sharpe Ratio. Kelly et al. (2021) explains that the use of centred (the traditional one) versus the uncentered is without loss of generality, as it does not change the qualitative conclusions drawn from those Sharpe Ratios.

Another important point is how they define the relationship between p, the number of parameters, and n, the number of time series observations, where $p/n = c$. When $c > 1$, it indicates an overparameterized model. Kelly et al. (2021) aims to test if the findings of Bartlett et al. (2020) and Tsigler and Bartlett (2023) apply to return prediction models.

Kelly et al. (2021) demonstrates that the largest possible approximating model should be used, meaning p should be much larger than n, as also shown by Bartlett et al. (2020). They observed that expected out-of-sample forecasting and portfolio performance improve with a higher model complexity.

2.2 R^2 for Judging Economic Benefits

R^2 is the coefficient of determination, a statistical measure that evaluates the quality of regression model predictions (James et al., 2023). R^2 represents the proportion of variance explained by the model. The formula is:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where

- $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*, where y_i is the observed value and $\bar{y}$ is the mean of the observed value.
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the *residual sum of squares*, where $\hat{y}_i$ is the value predicted by the model.

R^2 typically has a value between 0 and 1, but it can also be negative if the model performs worse at predicting than a simple horizontal line. In Kelly et al. (2021) paper, many of the models exhibit large negative R^2 values, even when the Sharpe Ratio is high. The authors also introduce a shrinkage z, which is useful for regularising the model. As explained in Section 1.2, regularisation helps to deal with high-dimensional data especially when the number of parameters exceeds the number of observations.

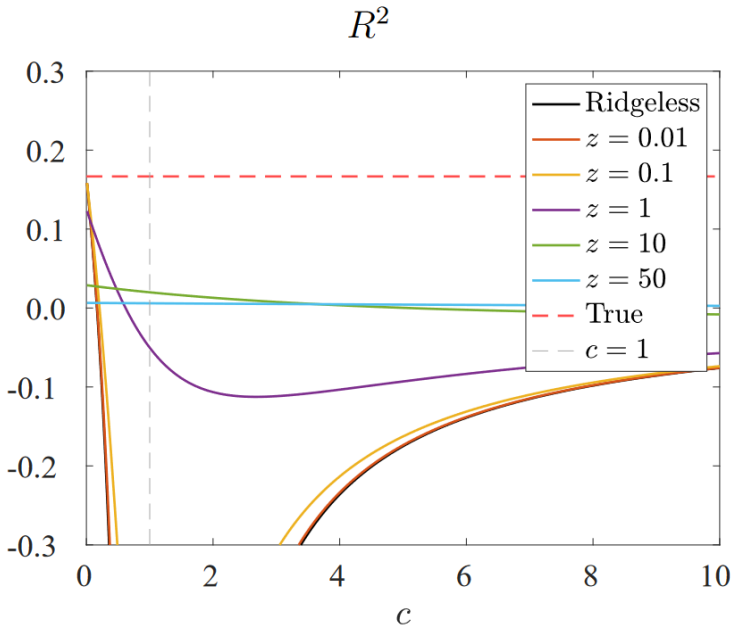


Figure I.4: Expected Out-of-sample R^2, (Kelly et al., 2021, p. 22).

In Figure I.4, we can observe several important points. First, when $c < 1$, the ridge parameter can improve R^2. In cases where $c = 1$, meaning the number of parameters equals the number of observations, R^2 can be highly negative, especially when the shrinkage is close to the ridgeless situation. However, when a large shrinkage is applied, the negativity of R^2 is reduced. Finally, when c is high, models with a shrinkage parameter z smaller than 1 show an improvement in R^2, but they still perform worse than models with a higher shrinkage parameter.

Introducing a positive ridge parameter pulls the coefficient estimates towards zero, which reduces the model's variance. This reduction often leads to an improvement in the out-of-sample R^2, particularly in complex models. This demonstrates the importance of regularisation in managing high-dimensional data and enhancing the predictive performance of regression models.

Kelly et al. (2021) indicates that using R^2 to judge the economic benefit of a machine learning model can be theoretically not useful, and that out-of-sample R^2 from a prediction model is an incomplete measure of its economic value. The reason is that R^2 is highly influenced by the variance of forecasts. A very low out-of-sample R^2 indicates a highly volatile timing strategy. However, the properties of least squares imply that the expected out-of-sample return of a timing strategy is always positive. Therefore, as long as the timing variance is not too high and R^2 is not too negative, the Sharpe Ratio can be substantial. Kelly et al. (2021) also states that selecting the correct shrinkage maximises the expected out-of-sample model performance.

Kelly et al. (2021) also explains that when $p = n$, the regression exactly fits the training data but performs poorly out-of-sample. However, in a high-complexity model regime, where the number of predictors exceeds the number of observations ($p > n$), the situation changes and the R^2 improve. Kelly et al. (2021) also shows that in this high-complexity regime, ridgeless regression can achieve accurate out-of-sample forecasts despite perfectly fitting the training data. All these findings align with the views of Bartlett et al. (2020); Tsigler and Bartlett (2023) regarding benign overfitting and those of Li, Su, and Sejdinovic (2021); Mikhail Belkin et al. (2019) concerning the double descent curve and the interpolation point, but applied within a financial context.

2.3 Models Specification

Model specification is essential for creating an effective model with a given dataset. To understand the differences between the correctly specified and the misspecified models, we need to define and explain their characteristics first (Kelly et al., 2021). This will enable us to evaluate the performance of each model in handling data, understand the implications for their predictive accuracy, assess their capacity to generalise, and potentially deal with benign overfitting.

Correctly Specified Models A correctly specified model should accurately capture the underlying function, so the relationship between predictors and the target variable should be well-represented. For these models, the amount of shrinkage that optimises the expected out-of-sample R^2 also optimises the Sharpe Ratio. This means that by fine-tuning the prediction model, we can ensure that the model performs well in both predictive accuracy and economic value.

Misspecified Models A misspecified model normally fails to accurately capture the true function, often due to oversimplification or incorrect assumptions. However, Kelly et al. (2021) shows that increasing model complexity improves out-of-sample market timing performance, even in cases where the models are not accurately specified.

Complexity is beneficial because it enables a better approximation of the true function, resulting in higher expected returns and a better Sharpe Ratio. Kelly et al. (2021) also explains that while smaller models are better when the model is correctly specified, misspecified models benefit significantly from increased complexity.

2.4 Double Ascent Curve

In the paper by Kelly et al. (2021), the authors discuss the discovery of the double ascent curve, which is related to the double descent curve but applied to the Sharpe Ratio. This curve is decomposed into three distinct phases and can be observed in Figure III.3.

First, we start by describing the ridgeless case where there is an increase in complexity. As complexity increases, measured by the ratio $p/n = c$, the Sharpe Ratio rises because the model captures more of the underlying signal. The second phase is the intermediate decline. After reaching a certain point, the Sharpe Ratio declines due to overfitting, because the model starts capturing noise instead of the signal. Finally, there is a second increase in the Sharpe Ratio after $c = 1$. With further complexity, especially in high-dimensional settings where $p >> n$, the Sharpe Ratio improves more than before. This counter-intuitive improvement is attributed to the model exploiting implicit regularisation and benign overfitting. In this phase, the model fits the training data perfectly but still generalises well.

Kelly et al. (2021) and Kelly et al. (2022) also demonstrate that increasing the amount of shrinkage makes complexity a "virtue" even in the low complexity regime (when $cq < 1$). In this scenario, the hump in the curve disappears, and the "double ascent" curve transitions into a "permanent ascent" curve. The parameter q represents the ratio of the number of latent factors (p_1) and the number of predictors (p), defined as $q = \frac{p_1}{p}$. Latent factors are hidden variables that may influence the observed data. This parameter essentially measures the relative proportion of the unobservable factors compared to the observable predictors used in the model. This double ascent curve is also observed by Didisheim et al. (2023).

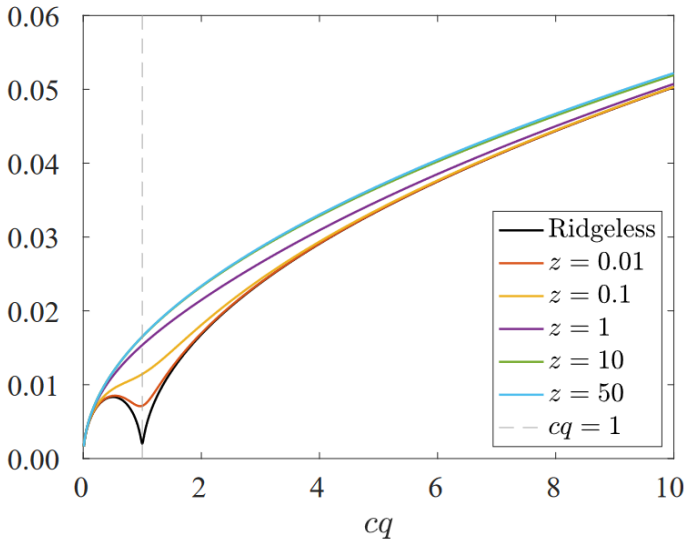


Figure I.5: Expected out-of-sample Shape Ratio from mis-specified models, (Kelly et al., 2021, p. 37).

3 Benign Overfitting in Portfolio Construction

3.1 Introduction to the Modern Portfolio Theory

Portfolio selection is an important process in investments that aims to maximise returns while managing risks. This involves carefully selecting a mix of different assets to create a balanced portfolio that aligns with the investor's risk tolerance, investment goals, and market conditions. The objective is to achieve optimal chosen performance metrics, ensuring that the portfolio is well-diversified and capable of resisting market fluctuations. In this section, we will examine four different portfolios that will be useful for our analysis of benign overfitting in modern portfolio theory.

3.2 Ledoit-Wolf Covariance Estimation

In this master's thesis, we work with large datasets where the number of assets p exceeds most of the time the number of observations n. Consequently, the covariance matrix is not directly invertible. To resolve this issue, we use the Ledoit-Wolf shrinkage estimator for the covariance matrix (Ledoit & Wolf, 2004). This shrinkage-based method is used to improve the covariance matrix estimation, which is important for constructing optimal portfolios. Traditional calculation of sample covariance matrices can be unstable and imprecise, particularly when the number of assets (p) is large compared to the sample size (n). The Ledoit-Wolf method avoids this problem by shrinking the sample covariance matrix towards a stable target matrix.

The Ledoit-Wolf shrinkage estimator of the covariance matrix $\boldsymbol{\Sigma_{LW}}$ is given by:

$$\Sigma_{LW} = \alpha \mathbf{T} + (1 - \alpha)\Sigma$$

where:

- $\boldsymbol{\Sigma}$ is the covariance matrix.
- $\mathbf{T}$ is the target matrix, which can be chosen based on the problem structure. A common choice for $\mathbf{T}$ is a scaled identity matrix, representing a model where all assets have the same variance and are uncorrelated.
- α is the shrinkage, this parameter can have a value between 0 and 1 that determines the weight given to the target matrix versus the covariance matrix.

Ledoit and Wolf (2004) proposed an optimal shrinkage intensity α that minimises the mean squared error (MSE) between the covariance matrix $\boldsymbol{\Sigma}$ and the estimated covariance matrix $\boldsymbol{\Sigma}_{LW}$.

3.3 Mean-Variance Portfolio

One of the most important frameworks in Modern Portfolio Theory is the mean-variance optimization model also called the Maximum Sharpe Ratio (MSR) introduced by Markowitz (1952). This model aims to maximise the expected return of a portfolio while simultaneously minimizing its risk which is mostly due to the variance of returns. This aims to maximise the Sharpe Ratio:

$$\max_{\mathbf{w}} \frac{\mathbf{w}'\boldsymbol{\mu}}{\sqrt{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}} \quad \text{subject to} \quad 1'\mathbf{w} = 1,$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean returns and covariance matrix of returns, respectively. $\mathbf{w}$ represents the portfolio weights for each possible asset.

The optimisation problem can be resolved by:

$$\hat{\mathbf{w}}_{MSR} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$$

The MSR portfolio has some issues because, most of the time, historical returns are not reliable for predicting future returns, as explained by Black (1993). Normally, this portfolio model does not yield the best out-of-sample Sharpe Ratio.

3.4 Minimum Variance Portfolio

Another well-known portfolio optimisation model is the Minimum Variance (MV) portfolio. This model aims to minimise the portfolio risk without considering the expected return.

The weights of the MV portfolio can be found by solving the following optimisation problem:

$$\min_{\mathbf{w}} \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \quad \text{subject to} \quad 1'\mathbf{w} = 1,$$

The optimal weights are given by the following formula:

$$\hat{\mathbf{w}}_{MV} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}$$

The minimum variance portfolio can have a better out-of-sample Sharpe Ratio than the mean-variance portfolio.

3.5 Naive Portfolio

The simplest existing portfolio is the Naive portfolio, also known as the equally weighted portfolio. This strategy involves diversifying by investing equal weights in all available assets (DeMiguel et al., 2009). This portfolio is defined as follows:

$$\hat{\mathbf{w}}_{naive} = \frac{1}{p}$$

where p is the number of available assets.

The Naive portfolio can perform well out-of-sample and is easy to implement. It incurs zero transaction costs and, sometimes, can outperform more complex portfolio strategies.

3.6 Bona-Fide Generalised Shrinkage Estimator

Bodnar et al. (2018) paper deals with the construction and evaluation of a shrinkage estimator (α) for a portfolio in a high-dimensional setting, where the number of assets p and the sample size n grows proportionally ($p/n = c$).

Shrinkage Estimator

The weights $\hat{\mathbf{w}}_{BFGSE}$ are obtained from the Bona-Fide Generalised Shrinkage Estimator (BFGSE):

$$\hat{\mathbf{w}}_{BFGSE} = \widehat{\alpha}^* \left(\frac{\mathbf{S}_n^+ \mathbf{1}}{\mathbf{1}' \mathbf{S}_n^+ \mathbf{1}} \right) + (1 - \widehat{\alpha}^*) \mathbf{b}_n$$

where $\mathbf{S}_n^+$ is the pseudo-inverse matrix of the sample covariance matrix $\mathbf{S}_n$. We use $\mathbf{S}_n^+ = \mathbf{S}_n^{-1}$ if $c < 1$.

The different variables are defined as follows:

- $\widehat{\alpha}^*$ is the optimal shrinkage intensity,
- $\mathbf{S}_n$ is the sample covariance matrix,
- $\mathbf{b}_n$ is the target portfolio, which can be either the Naive portfolio or the minimum variance portfolio.

Optimal Shrinkage Intensity ($\widehat{\alpha}^*$)

The optimal shrinkage intensity $\widehat{\alpha}^*$ in the case $c < 1$ is given by:

$$\widehat{\alpha}^* = \frac{(1 - \frac{p}{n})\hat{R}_{bn}}{\frac{p}{n} + (1 - \frac{p}{n})\hat{R}_{bn}}$$

For $c \geq 1$, the optimal shrinkage intensity $\widehat{\alpha}^*$ is:

$$\widehat{\alpha}^* = \frac{(\frac{p}{n} - 1)\hat{R}_{bn}}{(\frac{p}{n} - 1)^2 + \frac{p}{n} + (\frac{p}{n} - 1)\hat{R}_{bn}}$$

Estimator $\hat{R}_{bn}$

For $c < 1$:

$$\hat{R}_{bn} = (1 - \frac{p}{n})\mathbf{b}_n'\mathbf{S}_n\mathbf{b}_n \cdot \mathbf{1}'\mathbf{S}_n^{-1}\mathbf{1} - 1$$

For $c \geq 1$:

$$\hat{R}_{bn} = \frac{p}{n}(\frac{p}{n} - 1)\mathbf{b}_n'\mathbf{S}_n\mathbf{b}_n \cdot \mathbf{1}'\mathbf{S}_n^{+}\mathbf{1} - 1$$

The studies by Bodnar et al. (2018); Bodnar et al. (2023) suggest that the performance of the Sharpe Ratio is significantly influenced by the ratio c. When $c < 1$, traditional portfolios, such as the Maximum Sharpe Ratio (MSR) and Minimum Variance (MV) portfolios, as well as the Naive portfolio, often outperform the Bona-fide shrinkage estimator, especially when the target portfolio is the Naive portfolio. However, when $c > 1$, the Bona-fide shrinkage estimator generally outperforms the other portfolios mentioned above. Thus, the Sharpe Ratio is better when employing the Bona-fide shrinkage estimator in high-dimensional settings.

4 Research Question

Benign overfitting is a complex phenomenon that challenges traditional statistical theories. Understanding when and under which circumstances benign overfitting occurs is still a subject of ongoing research. While most studies focus on image datasets to analyse benign overfitting, this master thesis aims to extend these observations to financial data.

Research Question:
Does benign overfitting manifest in financial time series within high-dimensional datasets and can it be used for portfolio optimisation?

Sub-Questions:

1. Do the conditions given by Bartlett et al. (2020) effectively allow the identification of benign overfitting on financial data?
2. Does the occurrence of the double descent curve phenomenon arise in financial time series?
3. How does shrinkage influence the double descent curve and overfitting?
4. Can high-dimensional models exhibiting benign overfitting improve portfolio performance and reduce volatility?
5. Does the double descent curve appear in the context of high-dimensional portfolio models?

Chapter II

Data & Methodology

1 Dataset Description

To conduct the analysis and answer the research questions, we will use the Kenneth R. French Data Library. We selected 12 distinct sets of data to construct a daily dataset, referred to as **FF526-daily**, and a monthly dataset, referred to as **FF526-monthly**. Each dataset contains the same number of variables, with $p = 526$. In contrast, the number of observations differs, with $n = 15312$ for **FF526-daily** and $n = 730$ for **FF526-monthly**. The returns are value-weighted on a daily and monthly basis from July 1963 to April 2024. Missing values will be replaced with zero. The risk-free rate (RF), located in column 2 of each dataset, will be subtracted from all variables except for the 'Mkt-RF' column, where the risk-free rate has already been deducted. Then, all the data is divided by 100 to convert the values from percentages to the actual value. The datasets utilised for this analysis are the following:

Datasets	**Assets (p)**
Fama/French 5 Factors - *Mkt-RF & RF columns*	2
100 Portfolios Formed on Size and Investment (10 x 10)	100
100 Portfolios Formed on Size and Operating Profitability (10 x 10)	100
100 Portfolios Formed on Size and Book-to-Market (10 x 10)	100
25 Portfolios Formed on Size and Book-to-Market (5 x 5)	25
25 Portfolios Formed on Size and Operating Profitability (5 x 5)	25
25 Portfolios Formed on Size and Investment (5 x 5)	25
25 Portfolios Formed on Book-to-Market and Operating Profitability (5 x 5)	25
25 Portfolios Formed on Book-to-Market and Investment (5 x 5)	25
25 Portfolios Formed on Operating Profitability and Investment (5 x 5)	25
49 Industry Portfolios	49
25 Portfolios Formed on Size and Short-Term Reversal (5 x 5)	25

Table II.1: List of Kenneth R. French datasets used for the analysis.

2 Methodology

The empirical study of this master's thesis is divided into three distinct sections. The first section examines the conditions and the existence of benign overfitting within regression models. The subsequent section investigates the double descent curve phenomenon. The final section explores the application of benign overfitting to modern portfolio theory on high-dimensional datasets.

2.1 Benign Overfitting on Financial Data

In the first part, we will examine the conditions for achieving benign overfitting as proposed by Bartlett et al. (2020). The conditions to be tested are the following:

- **Condition 1**: $p > n$
  This condition ensures that the number of variables (p) exceeds the number of observations (n), which is an important requirement for overparameterization.
- **Condition 2**: $R_k(\Sigma) \gg n$
  $R_k(\Sigma)$ is the effective rank of the covariance matrix Σ. This condition implies that $R_k(\Sigma)$ should be much larger than the number of observations, ensuring that there are many directions in the space with a small variance. $R_k(\Sigma)$ formula can be found in Chapter I, Section 1.1.
- **Condition 3**: Dominant first eigenvalue
  This condition states that the first eigenvalue of Σ should be significantly larger than the other eigenvalues, indicating a dominant direction in the data's variance structure.
- **Condition 4**: $\sum_{i=1}^{p} \lambda_i \ll n$
  The sum of the eigenvalues of the covariance matrix should be much smaller than the number of observations by ensuring that the scale of the problem is small compared to the sample size.

We will first evaluate whether the selected datasets satisfy these conditions. Then, we will apply a multiple linear regression model to the dataset. The market returns (Mkt-RF) will be used as the outcome vector $y \in \mathbb{R}^n$. The other variables except the risk-free rate (RF), will be used as the data matrix $X \in \mathbb{R}^{n \times p_m}$ ($p_m = 524$). We will modify the dataset by selecting different numbers of observations for testing the regressions, with $n = [100;\ 200;\ 300;\ 400;\ 500]$. Two selection modes will be applied. Observations will be taken either from the "first" part of the dataset (starting in July 1970) or from the "last" part of the dataset (ending in April 2024) to compare the results. The dataset will be split into a training set (80%) and a test set (20%). To determine if benign overfitting

is occurring, we will assess whether the excess risk $R(\theta)$ is close to zero. The complete definition of $R(\theta)$ can be found in Chapter I, Section 1.1. To facilitate the calculation of $R(\theta)$, we will assume that we subtract the training MSE from the test MSE, given that the training MSE should be very small.

We will also verify the condition proposed by Mallinar et al. (2022); Mallinar et al. (2024) explained in Chapter I, Section 1.4 to classify the exact type of overfitting and define the category to which our financial datasets could belong.

Lastly, we will use ridge regression and introduce different shrinkage parameters (λ) to examine the influence of shrinkage on overparameterized models. The shrinkage parameters used will be $\lambda = [0;\ 1 \times 10^{-6};\ 1 \times 10^{-5};\ 1 \times 10^{-4};\ 0.001,\ 0.01;\ 0.1]$.

2.2 Double Descent Curve

In the second part of the research, we will use the previous results observed on regression models, starting with the multiple linear regression model, to investigate the double descent curve and analyse the risk so the mean squared error (MSE) for different numbers of variables (p) (changing between 5 and 524) while keeping the number of observations (n) constant. We will test this for $n = [100;\ 200;\ 300;\ 400;\ 500;\ 655;\ 700]$. This test aims to check if the double descent exists and how it manifests in the context of our financial datasets. This test is important for determining the optimal number of variables needed in a misspecified model, especially when the significance of each variable in the model is uncertain.

We will analyse more precisely these three situations:

- The MSE before the interpolation point ($p < n$).
- The MSE at the interpolation point ($p = n$).
- The MSE after the interpolation point ($p > n$).

In the second phase, we will decompose the test MSE into its two main components: the variance and the bias squared. This decomposition will help us to evaluate the impact of each MSE component on the potential double descent curve. We will implement the variance and the bias for a fixed n while continuously changing p to visualise the effects. The decomposition of the MSE is as follows:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}, \theta)$$

At last, we will also utilise the ridge regression model with the previously set shrinkage parameters λ. It will help us to examine the impact of λ on the double descent curve.

2.3 Portfolio Construction and Overfitting

In this last section, we will investigate whether benign overfitting exists within modern portfolio theory by comparing portfolio optimisation strategies. The models under consideration will be the Maximum Sharpe Ratio (MSR) portfolio, the minimum variance (MV) portfolio, the Naive portfolio and the optimisation portfolio based on the Bona-Fide Generalised Shrinkage Estimator (BFGSE). These models are detailed in Chapter I, Section 3. We will use the notation P to describe the different portfolios in the mathematical equations. The strategies will be based on the monthly dataset and will be long only to reduce the risk of having really small variances due to the dataset dimension. We will compare these models under three conditions: $c < 1$, $c = 1$, and $c > 1$. The specific values of c are $[0.75;\ 1;\ 1.5]$, where $p/n = c$ and p always contain the 524 variables (Without the Mkt-RF and the RF columns). The portfolio's performances will be evaluated using a 6-months rolling window for out-of-sample data and assessed across three metrics:

1. **Annual Average Returns:**

$$R^P_{\text{annual}} = \left(\frac{1}{n_w} \sum_{i=1}^{n_w} R^P_{i,\ \text{monthly}} \right) \times 12$$

   Where n_w is the number of months for the rolling window.

2. **Annual Volatility:**

$$\sigma^P_{\text{annual}} = \sqrt{\left(\frac{1}{n_w} \sum_{i=1}^{n_w} VAR^P_{\text{i},\ \text{monthly}} \right) \times 12}$$

3. **Sharpe Ratio (SR):**

$$\text{Sharpe Ratio} = \frac{R^P_{annual}}{\sigma^P_{annual}}$$

Lastly, the most important part will be to examine whether there is a variation in portfolio variance when the number of p changes while n remains constant. We will determine if there is evidence of a double descent and understand the consequences when c rises on the portfolio's variances. This will help us to assess if there is any indication of potential benign overfitting.

Chapter III

Results & Analysis

1 Results Presentation

In this section, we start by presenting the various results obtained from different parts of the study.

1.1 Benign Overfitting on Regression Models

To start this section, we will evaluate the results obtained from the **FF526-monthly** and **FF526-daily** datasets. Only the monthly dataset results will be detailed. Nevertheless, some daily dataset results are available in the Appendix. A general observation that we can give is that both datasets yielded the same results concerning the benign overfitting conditions. The first condition where p should be larger than n is pre-established by our settings and, therefore, does not require further verification. In our tests, the only scenarios that do not meet this condition are when $n = 700$ and when $c = 1$. For other cases, p is always larger than n and the condition 1 is satisfied.

However, the second condition concerning the effective rank is never met. The value of $R_k(\Sigma)$ is always smaller than n. This is due to the eigenvalues of the covariance matrix. On the other hand, the third condition is satisfied across all different n, as there is always a significantly large first eigenvalue compared to the second one. Similarly, the fourth condition is always met, with the sum of the eigenvalues being significantly smaller than n.

In conclusion, with these datasets, we verified three out of the four conditions. This indicates that we could achieve benign overfitting. The detailed observations can be found in Tables III.1, III.2. If a condition is verified, it is indicated as TRUE. If not, it is indicated as FALSE.

To understand the results of the dataset condition and the MSE in the context of possible benign overfitting, as well as the results on the double descent curve, we will introduce the

parameter n_{train}. As explained in the methodology, we split the dataset into training and testing parts. The training part constitutes 80% of the total dataset, so $n_{train} = 0.8 \times n$. This is important because the interpolation point of our results will be when $n_{train} = p$. We also define that $c = p/n_{train}$. All the observations and conditions for benign overfitting in the monthly dataset can be found in the table III.1.

n	**100**	**200**	**300**	**400**	**500**	**600**	**655**	**700**
n_{train}	80	160	240	320	400	480	524	560
condition 1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
c	6.55	3.27	2.18	1.64	1.31	1.09	1	0.94
condition 2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
$R_k(\Sigma)$	61.1011	101.4975	133.1122	156.2970	170.9266	183.9540	190.5702	194.2840
k **optimal**	11	27	37	52	60	66	66	71
condition 3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$1^{st}\lambda$	1.3182	1.5964	1.5900	1.3718	1.3861	1.4304	1.3672	1.4265
condition 4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$\sum \lambda_p$	1.7399	2.0143	1.9989	1.7648	1.8787	1.9142	1.8447	1.9277
MSE train	1.52E-31	2.85E-32	1.96E-31	1.01E-32	9.94E-32	3.39E-32	2.25E-32	9.58E-08
MSE test	2.54E-06	3.25E-06	7.03E-06	8.56E-06	2.24E-05	3.77E-05	5.25E-04	5.03E-05
$R(\theta)$	2.54E-06	3.25E-06	7.03E-06	8.56E-06	2.24E-05	3.77E-05	5.25E-04	5.02E-05

Table III.1: Benign overfitting conditions and MSE for $p = 524$, selection mode = "first", **FF526-monthly.csv**.

Regarding the Mean Squared Error and particularly the excess risk $R(\theta)$, we observe that when $p > n_{\text{train}}$, the train MSE is zero, indicating overfitting and an overtrained model. However, a significant finding is that the test MSE is lower when $p > n_{\text{train}}$ compared to when $p < n_{\text{train}}$, suggesting a reduced risk when the model is overfitted. We can also note that $R(\theta)$ is smaller when $c > 1$.

When we set $n_{\text{train}} = p$, we still achieve a zero training MSE. Nevertheless, by examining the test MSE, we notice that it is relatively higher at this interpolation point compared to when n is either larger or smaller than p. This observation raises the possibility of a double descent curve.

In Table III.2, we observe similar results for the selection mode "last", where the higher test MSE is when $c = 1$. This indicates that something significant occurs at $c = 1$, which we will analyse in detail in this Chapter, Section 1.2. The results for the daily dataset, which led to similar observations, are available in Appendix A.1; A.2.

n	**100**	**200**	**300**	**400**	**500**	**600**	**655**	**700**
n_{train}	80	160	240	320	400	480	524	560
condition 1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
c	6.55	3.27	2.18	1.64	1.31	1.09	1	0.94
condition 2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
$R_k(\Sigma)$	54.956	95.434	124.692	148.821	167.569	182.041	188.917	193.807
k **optimal**	17	38	59	65	71	73	71	73
condition 3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$1^{st}\lambda$	1.830	1.789	1.559	1.383	1.391	1.459	1.503	1.495
condition 4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$\sum \lambda_p$	2.505	2.356	2.208	1.975	1.935	1.980	2.023	2.013
MSE train	9.56E-32	1.99E-31	4.30E-33	9.87E-33	6.19E-32	8.76E-33	8.85E-32	1.01E-07
MSE test	9.54E-06	5.41E-06	1.80E-05	1.46E-05	2.30E-05	4.38E-05	0.000326	3.18E-05
$R(\theta)$	9.54E-06	5.41E-06	1.80E-05	1.46E-05	2.30E-05	4.38E-05	0.000326	3.17E-05

Table III.2: Benign overfitting conditions and MSE for $p = 524$, selection mode = "last", **FF526-monthly.csv**.

Concerning the ridge regression results, we tested the different values of shrinkage λ detailed in Chapter II, Section 2.1. When $\lambda = 0$, we are in the case of ridgeless regression, resulting in the same outcomes as those obtained in linear regression (Table III.1, III.2). We observed that when $p > n_{\text{train}}$, shrinkage is not particularly useful when c is very high. For example, as shown in Table III.3, at $c = 6.55$, increasing λ results in both the test and train MSE becoming worse.

λ	**0**	**1E-06**	**1E-05**	**0.0001**	**0.001**	**0.01**	**0.1**
train MSE	4.96E-31	1.41E-16	1.41E-14	1.41E-12	1.40E-10	1.26E-08	6.66E-07
test MSE	2.54E-06	2.54E-06	2.54E-06	2.54E-06	2.55E-06	2.65E-06	4.06E-06
$R(\theta)$	2.54E-06	2.54E-06	2.54E-06	2.54E-06	2.55E-06	2.64E-06	3.40E-06

Table III.3: Ridge regression with different λ for $n = 100$ and $p = 524$, selection mode = "first".

However, when c is smaller ($c = 2.18$) but still higher than one, shrinkage starts to become useful and reduces the test MSE. For the shrinkage to be effective, the value of λ must be sufficiently high. If the shrinkage is too low, there is no significant change in the test MSE. The results, presented in Table III.4, show a change in the test MSE only when $\lambda = 1 \times 10^{-4}$.

λ	**0**	**1E-06**	**1E-05**	**0.0001**	**0.001**	**0.01**	**0.1**
train MSE	1.89E-30	1.71E-15	1.70E-13	1.69E-11	1.52E-09	7.35E-08	9.19E-07
test MSE	7.03E-06	7.03E-06	7.03E-06	7.01E-06	6.84E-06	5.91E-06	4.62E-06
$R(\theta)$	7.03E-06	7.03E-06	7.03E-06	7.01E-06	6.84E-06	5.83E-06	3.70E-06

Table III.4: Ridge regression with different λ for $n = 300$ and $p = 524$, selection mode = "first".

The main discovery here is when $c = 1$, we observe a significant reduction in the test MSE even for small values of λ. This indicates that at the interpolation point, where there is a substantial increase in the test MSE, applying shrinkage can highly influence the test MSE and the excess risk $R(\theta)$.

λ	**0**	**1E-06**	**1E-05**	**0.0001**	**0.001**	**0.01**	**0.1**
train MSE	8.51E-32	4.53E-09	1.89E-08	6.70E-08	1.89E-07	5.36E-07	1.55E-06
test MSE	5.25E-04	1.37E-04	6.20E-05	2.89E-05	1.91E-05	1.13E-05	8.69E-06
$R(\theta)$	5.25E-04	1.37E-04	6.19E-05	2.88E-05	1.89E-05	1.08E-05	7.14E-06

Table III.5: Ridge regression with different λ for $n = 655$ and $p = 524$, selection mode = "first".

All the previous observations are illustrated in Figure III.1. In this figure, it is clear that when c is very high, applying shrinkage results in worse test and train MSE values. However, when c is moderate, a higher shrinkage parameter leads to a slight reduction in the test MSE. Finally, when $c = 1$, there is a significant reduction in the test MSE as the shrinkage parameter increases. This demonstrates the impact of shrinkage across different values of c. The results for the daily dataset, using the same shrinkage parameters λ, are available in Appendix A.3.

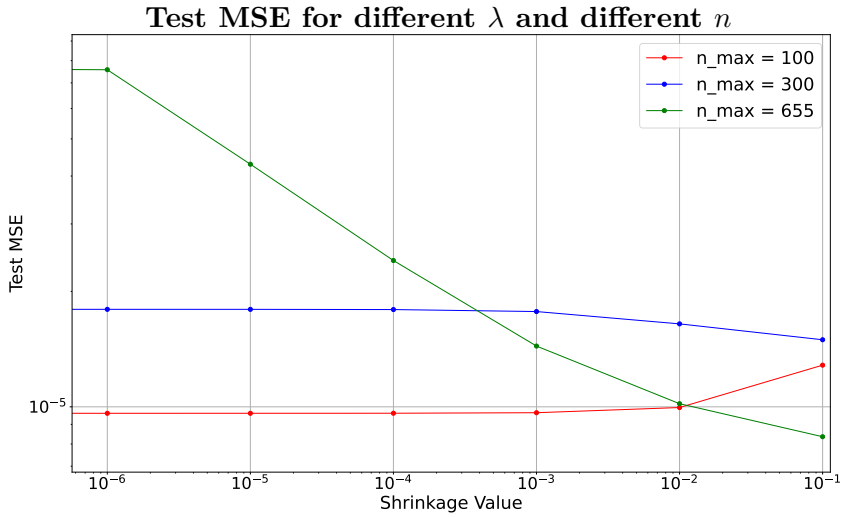


Figure III.1: Ridge regression test MSE with different shrinkage λ.

1.2 Double Descent Curve

Regarding the double descent curve, we tested various values of n and consistently obtained similar results to the previous section. When $c < 1$, the train MSE decreased continuously with the addition of more p, resulting in a continuous descending train MSE curve. At the interpolation point, the training MSE is zero and remains zero for all $p > n_{train}$. For the test MSE, it remained approximately constant until approaching the interpolation point

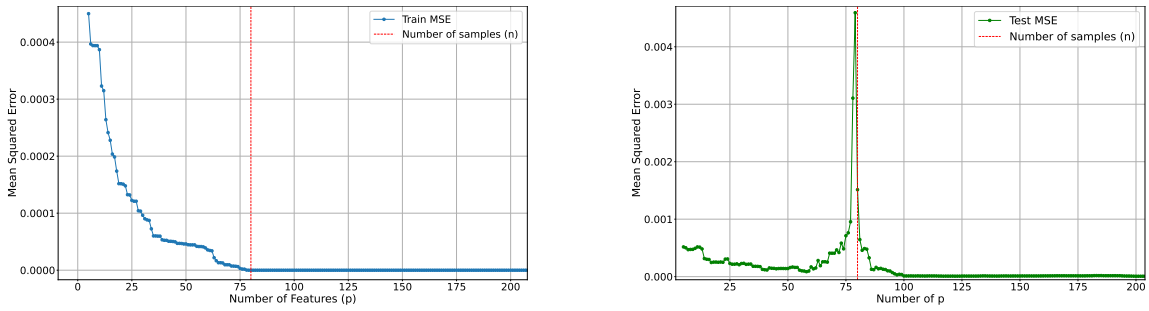
($n_{train} = p$), at which the test MSE spiked. However, the maximal test MSE did not occur exactly at the interpolation point, but it was more observed at $n - 1$ or $n - 2$ relative to the theoretical interpolation point. What is particularly noteworthy is that when $c > 1$, the resulting test MSE is consistently lower than it was before the interpolation point. This was observed across several different n values and in various tests on both the daily and monthly datasets.

In Table III.6, we can observe the test outcomes with varying n values, the table shows the maximum and minimum test Mean Squared Error (MSE) observed on the **FF526-monthly** dataset. The maximum test MSE consistently occurs when the parameter p equals the training sample size n_{train}. In contrast, the minimum test MSE is always observed when p exceeds n_{train}.

n	n_{train}	**minimum test MSE**	**optimal** p	**maximal test MSE**	**worst** p
100	80	9.35E-06	455	0.01196	79
200	160	5.41E-06	524	0.03057	159
300	240	1.52E-05	476	0.00158	238
400	320	1.37E-05	521	3.83995	319

Table III.6: Maximum and minimum test MSE results.

The situation when $n = 100$ and $n_{\text{train}} = 80$ is illustrated in Figure III.2. In this figure, the train MSE is represented in blue, showing that the MSE drops to zero when p is equal to n_{train} and remains at zero for all $p > n$. This indicates that it fits the training data perfectly. On the other hand, the test MSE is depicted in green. A spike is observed when p equals $n_{\text{train}} - 1$. After this spike, the test MSE begins to decrease, showing lower values compared to when n was larger than p. However, in this configuration, it is not possible to determine if the test MSE starts rising again when $p = n^2$, as explained by Li, Su, and Sejdinovic (2021).



(a) Train MSE vs number of p

(b) Test MSE vs number of p

Figure III.2: Double descent curve for different p and $n = 100$.

We also decided to develop a code to investigate whether the order and possible arrangement of the dataset influenced the double descent curve. To do this, we shuffled the columns of the dataset to create other orders and check if the double descent curve persisted. In each case, we observed that there was always a spike in the test MSE at the interpolation point.

As explained in the methodology, we decided to examine the variance and bias components of the test MSE. Our results indicate that the spike in the test MSE is primarily due to a significant increase in variance near the interpolation point. However, there is also a spike in the bias at that point, though it is relatively low compared to the increase in variance. Additionally, the scope of our investigations includes testing whether this kind of result will be obtained in the portfolio models when $c = 1$.

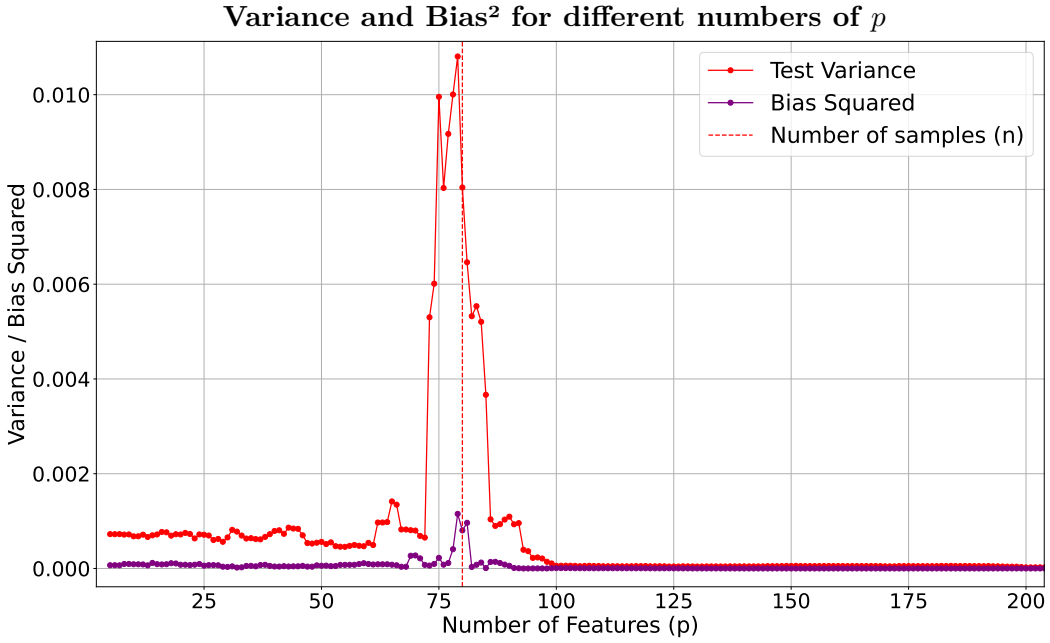


Figure III.3: Test MSE decomposition in variance and Bias² for different p and $n = 100$.

Considering ridge regression and the discussions in Chapter III, Section 1.1, we observed a significant reduction in the test MSE when the shrinkage parameter λ is high, particularly at the interpolation point. In Figure III.4, it is evident that higher λ values effectively truncate the peak of the curve, reducing the MSE at and slightly beyond the interpolation point. However, beyond the interpolation point, the curves for different λ values converge, with $\lambda = 0$ ultimately resulting in the lowest MSE when p is much larger than n. The results for this test on the daily dataset, which demonstrate the same results discussed in this section, are available in Appendix A.1.

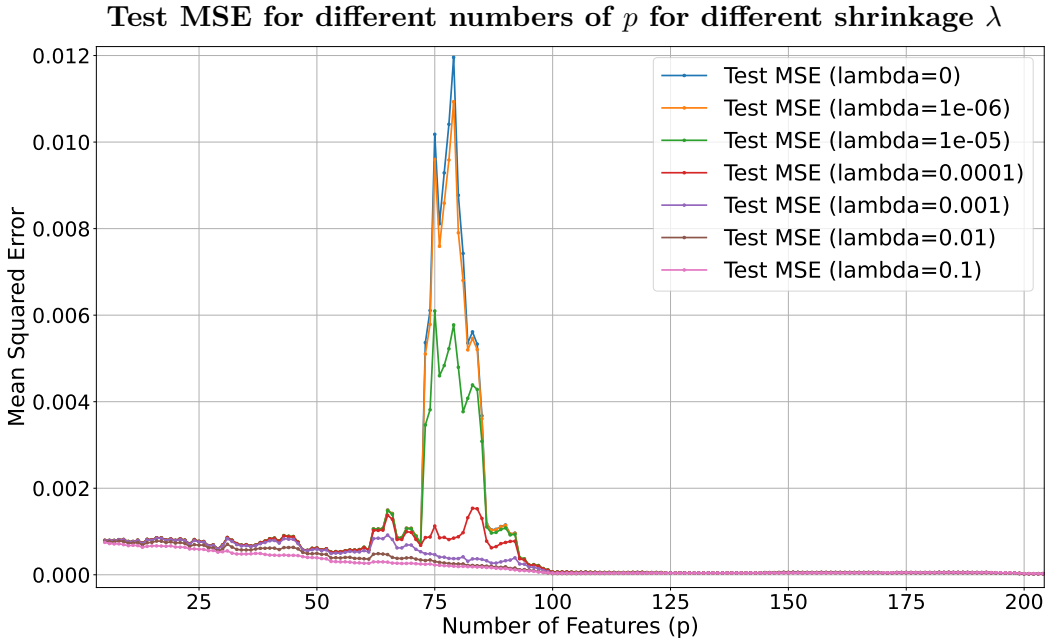


Figure III.4: Double descent curve in ridge regression with different shrinkages λ for a fixed $n = 100$.

1.3 Portfolio Theory and Overfitting

In this section, we will examine the behaviour of different portfolios for the values of $c = [0.75;\ 1;\ 1.5]$, with the number of p held constant at 524. The results are as follows:

Annual Performance Metrics			
	c = 0.75		
Strategy	**Returns**	**Volatility**	**SR**
MSR	0.005023	0.177399	0.028315
MV	-0.022919	0.147755	-0.155112
Naive	0.095268	0.240049	0.396870
BFGSE	0.093748	0.240952	0.389074
	c = 1		
Strategy	**Returns**	**Volatility**	**SR**
MSR	0.015826	0.164100	0.096441
MV	-0.046736	0.159526	-0.292967
Naive	0.095268	0.240049	0.396870
BFGSE	0.095268	0.240049	0.396870
	c = 1.5		
Strategy	**Returns**	**Volatility**	**SR**
MSR	0.052076	0.154905	0.336176
MV	-0.023905	0.158838	-0.150502
Naive	0.095268	0.240049	0.396870
BFGSE	0.093654	0.228707	0.409494

Table III.7: Annual performance metrics of portfolios for different values of c.

The Table III.7 indicates that, across all scenarios with varying c, the BFGSE portfolio has not always the highest Sharpe Ratio. The Naive portfolio maintains a constant value for all c, which is anticipated since the rolling window is fixed at 6 months, ensuring that the returns for each p remain unchanged regardless of c. This stability allows us to effectively compare the performance of other models as c varies.

Notably, as c increases, the Sharpe Ratio of the BFGSE model continues to improve, delivering better results despite its higher volatility compared to the MSR and MV models. In contrast, the minimum variance portfolio does not perform well in any scenario. Although it exhibits the lowest out-of-sample volatility even as c increases. Nevertheless, its returns are consistently negative, resulting in a negative Sharpe Ratio. Finally, the MSR portfolio shows enhanced performance with increasing c, reflecting an improvement in its Sharpe Ratio.

Ultimately, we investigate whether there is a reduction in portfolio variance as the number of p increases, while n remains constant. It is important to note the continuous decrease in portfolio variance as c increases in the MSR, MV, and Naive optimization models. However, in the BFGSE portfolio, variance rises near $c = 1$ due to shrinkage $\hat{\alpha}^*$ being equal to zero at that point, indicating that this is not a double descent curve. This investigation demonstrates a reduction in variance when $c > 1$, as illustrated in Figure III.5. To conclude, increasing the number of p generally leads to a reduction in out-of-sample portfolio variance.

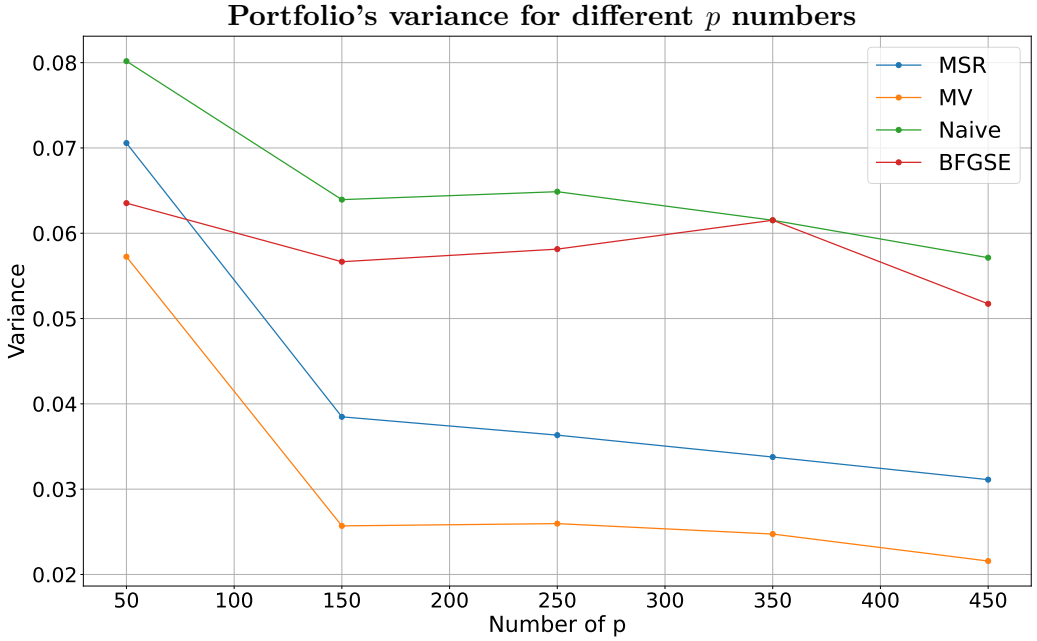


Figure III.5: Portfolio variance for different p and a fixed $n = 350$, $c = 0.75$

2 Results Analysis

Concerning the used datasets, we can say that they meet three out of the four conditions necessary for establishing benign overfitting given by Bartlett et al. (2020). Additionally, by examining the excess risk $R(\theta)$, we observe that the risk decreases as c increases, and it performs better than when $c < 1$ in every case. This suggests that we might be operating in a regime of benign overfitting or at least in a tempered overfitting scenario close to benign, as described by the taxonomy from Mallinar et al. (2022).

Furthermore, the results obtained from the double descent curve are also really significant. They indicate that the test MSE improves when p is larger than n, implying that the overfitting in our multiple linear regression model could be benign. This is especially notable because the minimum test MSE when $p > n$ is always better than the minimum test MSE when $p < n$. This suggests that when there is uncertainty about which variables are determinants, including more variables than observations is preferable. This double descent curve reinforces our belief that we are observing a case of benign overfitting.

Regarding the shrinkage parameter λ for ridge regression, we can observe that it can be useful in certain situations. As explained by Kelly et al. (2021) and Kelly and Xiu (2023), the amount of shrinkage can make complexity a virtue. Specifically, when we are close to the interpolation point ($c = 1$), where there is a clear reduction in the test MSE. However, it does not work in all cases, particularly when p is significantly larger than n. In such scenarios, if we are experiencing benign overfitting, the test MSE with shrinkage can sometimes yield worse results than in the ridgeless case ($\lambda = 0$) as shown in the results for $c = 6.55$ in Table III.3 and Figure III.1.

Regarding the MSE decomposition, we can conclude that the variance primarily drives the double descent curve, while the bias remains low when p is larger than n. Near the interpolation point, both bias and variance increase, indicating that the results at this point are not stable and that a dataset with $n = p$ should be avoided. Even if we apply shrinkage, as shown in Figure III.1, the test MSE will still be higher than when $p \gg n$.

Concerning the portfolio optimization models, we can note that the best model in terms of Sharpe Ratio is consistently the BFGSE portfolio when $c > 1$. This is not surprising, as this model is designed to work well with high-dimensional datasets. However, it is important to highlight the significance of the decreasing variance of the portfolios as p increases while n remains fixed. Despite this observation, it is not possible to definitively conclude that we are experiencing benign overfitting. This is because the results we obtained do not align with those observed in the multiple linear regression and ridge regression cases. Therefore, further analysis is needed to determine the nature of the overfitting in the portfolio optimisation models.

In conclusion, while the data must generally meet most conditions for benign overfitting, it is not the only influencing factor. The variations between models also play a significant role, as benign overfitting can be more evident in some models than in others. Furthermore, we can not guarantee that the conditions on datasets for benign overfitting are applicable to models other than regression models.

Conclusion

This master's thesis has analysed the phenomenon of benign overfitting, which has been predominantly studied using image recognition models, within the context of financial data. The primary objective was to determine whether benign overfitting can be applied in finance and modern portfolio theory.

To investigate the presence of benign overfitting in our selected financial data, we started by testing the benign overfitting conditions outlined by Bartlett et al. (2020). After verifying that 3 out of the 4 conditions were met, we calculated the excess risk $R(\theta)$ in a multiple linear regression model across the different datasets for $c > 1$, $c = 1$, and $c < 1$. We observed that $R(\theta)$ was lower when c was higher than one, and it further decreased as c increased. The excess risk was so low when $p > n$ that, according to the taxonomy established by Mallinar et al. (2022), we could conclude that we were in a regime of benign overfitting or at least tempered overfitting. Additionally, the train MSE remained zero when $p = n$, and at that point, the test MSE was at its maximum. This gave us the intuition that there might be a double descent curve for the linear regression model. As explained by Mikhail Belkin et al. (2019), the double descent curve can justify the presence of being in a regime of benign overfitting. To further investigate, we set the linear regression model with a fixed number of observations n and a varying number of variables p. This approach revealed the double descent curve with a noticeable spike at the interpolation point when $n = p$. The main observation is that the test MSE after the interpolation point was better than the ones observed at $c < 1$, confirming that we could be in a situation of benign overfitting under those circumstances.

We also applied a ridge regression model to examine the impact of the shrinkage parameter λ and compared it to the ridgeless case. We observed that the effectiveness of shrinkage depends on the value of c. Specifically, when c is very large, the test MSE results are worse than in the ridgeless case, making shrinkage unhelpful. However, near the interpolation point ($c = 1$), even a small shrinkage can be beneficial and significantly reduce the test MSE. This phenomenon can be observed on the double descent curve for the ridge regression model. Additionally, there are cases where the test MSE is reduced but only with a large λ, indicating that for some values of c, the test MSE remains stable across all λ. This

suggests that there is a critical value of c where the utility of shrinkage is neutral and at that point, the shrinkage may change from beneficial to detrimental for the test MSE. So we can state that a ridge regression with shrinkage can be useful, but only under the conditions mentioned above.

We also noticed that the spike in test MSE at the interpolation point is mainly due to an increase in variance, along with a rise in bias. However, after the interpolation point, the bias stays very low, and the variance decreases. We considered using this variance curve to check for potential variance increases in portfolio optimization when $c = 1$ and to explore the presence of a possible double descent curve in high-dimensional portfolio models.

For the portfolio optimisation section, we applied four different optimization models. Three of these are well-known: the Maximum Sharpe Ratio (MSR) portfolio, the Minimum Variance (MV) portfolio and the Naive portfolio. The fourth model, less famous but developed to perform better in high-dimensional settings when $p > n$, is the Bona-fide generalised Shrinkage Estimator (BFGSE) portfolio. We implemented these portfolios for three different values of $c = [0.75;\ 1;\ 1.5]$ and evaluated the models over a 6-month horizon using a rolling window approach. We observed that the BFGSE portfolio consistently yielded the highest Sharpe ratio when $c > 1$. Additionally, we demonstrated that for a fixed number of observations n and a varying number of variables p, the out-of-sample variance of the portfolio decreased as p increased. However, we did not observe any double descent curve or an increase in out-of-sample variance at the interpolation point. Consequently, we could not conclude that we were in a regime of benign overfitting for these portfolio models.

In conclusion, our findings indicate that benign overfitting can occur within financial data, driven by both the dataset and the model employed. If either the dataset or the model fails to meet the necessary conditions, the presence of benign overfitting cannot be confirmed. Additionally, the applicability of the benign overfitting conditions to models other than regression could maybe not be extended beyond regression models.

Bibliography

Ali, A., Kolter, J. Z., & Tibshirani, R. J. (2019). A Continuous-Time View of Early Stopping for Least Squares Regression. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 1370–1378. https://proceedings.mlr.press/v89/ali19a.html

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070. https://doi.org/10.1073/pnas.1907378117

Belkin, M., Hsu, D. J., & Mitra, P. (2018). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/e22312179bf43e61576081a2f250f845-Abstract.html

Belkin, M., Ma, S., & Mandal, S. (2018). To Understand Deep Learning We Need to Understand Kernel Learning. *Proceedings of the 35th International Conference on Machine Learning*, 541–549. https://proceedings.mlr.press/v80/belkin18a.html

Belkin, M., Rakhlin, A., & Tsybakov, A. B. (2019). Does data interpolation contradict statistical optimality? *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, *89*, 1611–1619. https://proceedings.mlr.press/v89/belkin19a.html

Black, F. (1993). Estimating Expected Return. *Financial Analysts Journal*, *49*(5), 36–38. https://doi.org/10.2469/faj.v49.n5.36

Bodnar, T., Okhrin, Y., & Parolya, N. (2023). Optimal shrinkage-based portfolio selection in high dimensions. *Journal of Business & Economic Statistics*, *41*(1), 140–156. https://doi.org/10.1080/07350015.2021.2004897

Bodnar, T., Parolya, N., & Schmid, W. (2018). Estimation of the Global Minimum Variance Portfolio in High Dimensions. *European Journal of Operational Research*, *266*(1), 371–390. https://doi.org/10.1016/j.ejor.2017.09.028

Chatterji, N. S., & Long, P. M. (2022). Foolish Crowds Support Benign Overfitting. *Journal of Machine Learning Research*, *24*(125), 1–12. http://jmlr.org/papers/v23/21-1199.html

Chen, J., Cao, Y., & Gu, Q. (2021). Benign Overfitting in Adversarially Robust Linear Classification. http://arxiv.org/abs/2112.15250

Chen, L., Lu, S., & Chen, T. (2022). Understanding Benign Overfitting in Gradient-Based Meta Learning. https://doi.org/10.48550/arXiv.2206.13482

DeMiguel, V., Garlappi, L., & Uppal, R. (2007). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, *22*(5), 1915–1953. https://doi.org/10.1093/rfs/hhm075

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/ N Portfolio Strategy? *Review of Financial Studies*, *22*(5), 1915–1953. https://doi.org/10.1093/rfs/hhm075

Devroye, L., Györfi, L., & Krzyżak, K. (1998). The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, *65*(2), 209–227. https://doi.org/https://doi.org/10.1006/jmva.1997.1725

Didisheim, A., Ke, S., Kelly, B. T., & Malamud, S. (2023). Complexity in Factor Pricing Models. *Swiss Finance Institute Research Paper No. 23-19*. https://doi.org/10.2139/ssrn.4388526

Fama, E. F., & French, K. R. (2023). Production of U.S. Rm-Rf, SMB, and HML in the Fama-French Data Library. *Chicago Booth Research Paper No. 23-22*. https://doi.org/10.2139/ssrn.4629613

Guan, X., & Burton, H. (2022). Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications. *Structures*, *46*, 17–30. https://doi.org/10.1016/j.istruc.2022.10.004

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, *50*(2), 949–986. https://doi.org/10.1214/21-AOS2133

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

Kelly, B. T., Malamud, S., & Zhou, K. (2021). The Virtue of Complexity in Return Prediction. *Swiss Finance Institute Research Paper No. 21-90, Journal of Finance, forthcoming*. https://doi.org/10.2139/ssrn.3984925

Kelly, B. T., Malamud, S., & Zhou, K. (2022). The Virtue of Complexity Everywhere. https://doi.org/10.2139/ssrn.4166368

Kelly, B. T., & Xiu, D. (2023). Financial Machine Learning. https://doi.org/10.2139/ssrn.4501707

Kenneth R. French - Data Library. (n.d.). https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Koehler, F., Zhou, L., Sutherland, D. J., & Srebro, N. (2021). Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting. *Advances in Neural Information Processing Systems*, *34*, 20657–20668. https://proceedings.neurips.cc/paper/2021/hash/ac9815bef801f58de83804bce86984ad-Abstract.html

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Li, Z., Su, W., & Sejdinovic, D. (2021). Benign Overfitting and Noisy Features. http://arxiv.org/abs/2008.02901

Li, Z., Zhou, Z.-H., & Gretton, A. (2021). Towards an Understanding of Benign Overfitting in Neural Networks. https://doi.org/10.48550/arXiv.2106.03212

Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., & Nakkiran, P. (2022). Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting. https://doi.org/10.48550/arXiv.2207.06569

Mallinar, N., Zane, A., Frei, S., & Yu, B. (2024). Minimum-Norm Interpolation Under Covariate Shift. http://arxiv.org/abs/2404.00522

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, *7*(1), 77–91. https://doi.org/10.2307/2975974

Mikhail Belkin, Belkin, M., Daniel Hsu, Hsu, D., Siyuan Ma, Ma, S., Soumik Mandal, & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0

Muthukumar, V., Vodrahalli, K., Subramanian, V., & Sahai, A. (2020). Harmless Interpolation of Noisy Data in Regression. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, *1*(1). https://doi.org/10.1109/JESTPE.2020.2985116

Peters, E., & Schuld, M. (2022). Generalization despite overfitting in quantum machine learning models. https://doi.org/10.48550/arXiv.2209.05523

Rakhlin, A., & Zhai, X. (2019). Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. *Proceedings of the Thirty-Second Conference on Learning Theory*, 2595–2623. https://proceedings.mlr.press/v99/rakhlin19a.html

Rice, L., Wong, E., & Kolter, J. Z. (2020). Overfitting in adversarially robust deep learning. https://doi.org/10.48550/arXiv.2002.11569

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., & Wyart, M. (2019). A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, *52*(47), 474001. https://doi.org/10.1088/1751-8121/ab4c8b

Tsigler, A., & Bartlett, P. L. (2023). Benign overfitting in ridge regression. *Journal of Machine Learning Research*, (123). http://jmlr.org/papers/v24/22-1398.html

Wibawa, A. P., Utama, A. B. P., Elmunsyah, H., Pujianto, U., Dwiyanto, F. A., & Hernandez, L. (2022). Time-series analysis with smoothed Convolutional Neural Network. *Journal of Big Data*, *9*(1), 44. https://doi.org/10.1186/s40537-022-00599-y

Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. https://arxiv.org/abs/2002.11328

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. http://arxiv.org/abs/1611.03530

Zhang, H., Jacobsen, B., & Jiang, F. (2019). Boosting Portfolio Choice in the Big Data Era. https://doi.org/10.2139/ssrn.3495901

Appendix

The results mentioned in this master's thesis for the daily dataset **FF526-daily.csv** are detailed below.

Linear Regression and Benign Overfitting Conditions

n	**100**	**200**	**300**	**400**	**500**	**600**	**655**	**700**
n_{train}	80	160	240	320	400	480	524	560
condition 1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
c	6.55	3.27	2.18	1.64	1.31	1.09	1	0.94
condition 2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
$R_k(\Sigma)$	60.19	95.4101	122.9005	143.5874	156.6956	169.8501	174.7016	177.9267
k **optimal**	9	24	35	41	45	49	51	50
condition 3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$1^{st}\lambda$	0.0083	0.01357	0.01127	0.0098	0.0104	0.0113	0.0111	0.0121
condition 4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$\sum \lambda_p$	0.02295	0.0303	0.0272	0.02531	0.02554	0.02679	0.02719	0.02868
MSE train	1.57E-33	2.80E-34	3.34E-35	3.99E-35	1.67E-34	1.44E-34	3.81E-33	3.59E-10
MSE test	1.27E-07	5.72E-08	6.25E-08	8.19E-08	9.16E-08	1.57E-07	1.62E-05	1.28E-07
$R(\theta)$	1.27E-07	5.72E-08	6.25E-08	8.19E-08	9.16E-08	1.57E-07	1.62E-05	1.27E-07

Table A.1: Benign overfitting conditions and MSE for $p = 524$, selection mode = "first".

n	**100**	**200**	**300**	**400**	**500**	**600**	**655**	**700**
n_{train}	80	160	240	320	400	480	524	560
condition 1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
c	6.55	3.27	2.18	1.64	1.31	1.09	1	0.94
condition 2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
$R_k(\Sigma)$	50.36	87.42	117.11	137.59	152.48	164.79	169.60	173.12
k **optimal**	21	41	61	67	74	79	79	77
condition 3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$1^{st}\lambda$	0.0589	0.0580	0.0614	0.0757	0.0918	0.0936	0.0910	0.0881
condition 4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
$\sum \lambda_p$	0.0849	0.0831	0.0874	0.1037	0.1223	0.1264	0.1240	0.1206
MSE train	3.29E-33	4.36E-34	3.26E-33	3.44E-33	4.39E-33	1.45E-34	8.95E-33	9.97E-10
MSE test	6.81E-08	4.83E-08	8.15E-08	1.47E-07	1.97E-07	3.32E-07	2.17E-06	3.30E-07
$R(\theta)$	6.81E-08	4.83E-08	8.15E-08	1.47E-07	1.97E-07	3.32E-07	2.17E-06	3.29E-07

Table A.2: Benign overfitting conditions and MSE for $p = 524$, selection mode = "last".

Ridge Regression and Shrinkage

c = 6.55							
λ	**0**	**1.00E-06**	**1.00E-05**	**0.0001**	**0.001**	**0.01**	**0.1**
train MSE	1.75E-36	1.75E-15	1.75E-13	1.72E-11	1.51E-09	7.57E-08	1.23E-06
test MSE	1.27E-07	1.27E-07	1.27E-07	1.28E-07	1.39E-07	2.56E-07	1.08E-06
$R(\theta)$	1.27E-07	1.27E-07	1.27E-07	1.28E-07	1.37E-07	1.81E-07	-1.50E-07
c = 2.18							
λ	**0**	**1.00E-06**	**1.00E-05**	**0.0001**	**0.001**	**0.01**	**0.1**
train MSE	7.27E-34	5.85E-15	5.70E-13	4.55E-11	1.55E-09	3.05E-08	5.03E-07
test MSE	6.24E-08	6.23E-08	6.18E-08	5.78E-08	4.70E-08	7.09E-08	3.53E-07
$R(\theta)$	6.24E-08	6.23E-08	6.18E-08	5.78E-08	4.54E-08	4.04E-08	-1.50E-07
c = 1							
λ	**0**	**1.00E-06**	**1.00E-05**	**0.0001**	**0.001**	**0.01**	**0.1**
train MSE	7.31E-33	1.96E-10	5.56E-10	1.48E-09	3.76E-09	1.99E-08	2.51E-07
test MSE	1.62E-05	6.15E-08	3.86E-08	3.52E-08	3.15E-08	5.33E-08	3.77E-07
$R(\theta)$	1.62E-05	6.13E-08	3.81E-08	3.37E-08	2.78E-08	3.34E-08	1.25E-07

Table A.3: Ridge regression with different α for different c, selection mode = "first".

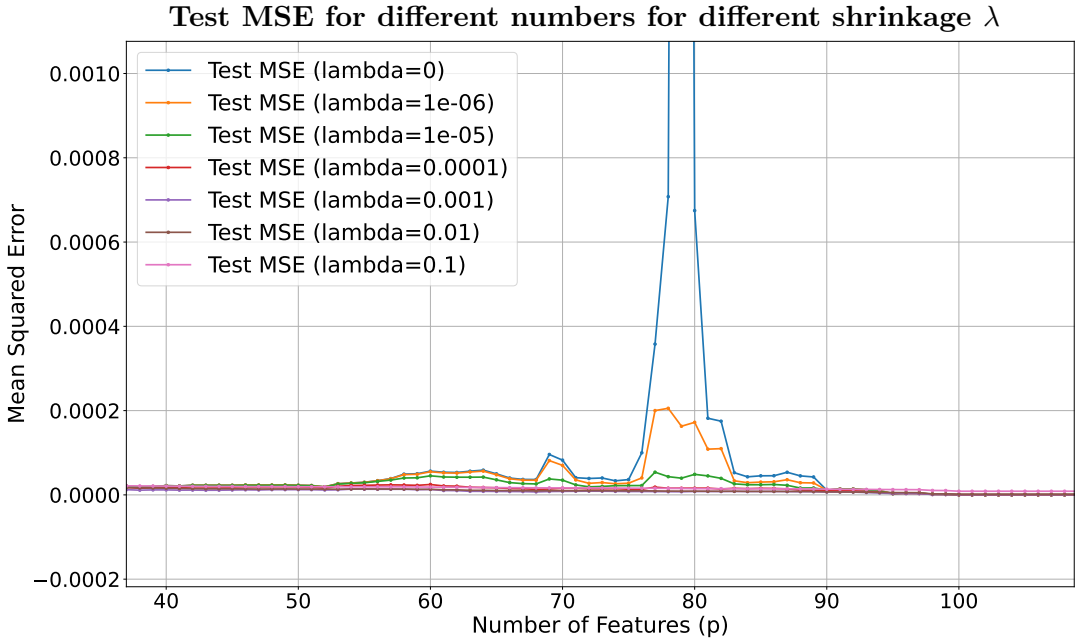


Figure A.1: Double descent curve in ridge regression with different shrinkages λ and a fixed $n = 100$.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**